DISPLAY AND INTERPRET DATA

This unit is about organizing and interpreting data. Stem-and-leaf plots may be used to order and organize data. Data can be interpreted by calculating measures of variation and organizing the data into box-and-whiskers plots. After processing data into an ordered display, predictions may be made based upon the trends of the data.

Stem-and-Leaf Plots

A way to organize data is by using a stem-and-leaf plot. The leaf is the last digit of a number. The stem is remaining digits of a number. Stems are listed in numerical order to create categories.

In the example shown below, the leaves is the ones digit of the numbers and the stems are first digits (tens digits).

Example: Arrange the following numbers in a stem-and-leaf plot: 58, 72, 53, 95, 62, 67, 58, 84, 77, 52, 89, 96, 64, 81.

First we'll make a rough draft of our stem-and-leaf plot.

first digits

stem		**leaf**		
5	8	3	8	2
6	2	7	4	
7	2	7		
8	4	9	1	
9	5	6		

ones digits

Key: 5 | 8 represents 58.

*The 5 in the stem column together with the 8 in the leaf column represent the number 58 from the data points.

Now, we'll rewrite the stem-and-leaf plot placing the leaves in order from left to right.

first digits

stem		leaf		
5	2	3	8	8
6	2	4	7	
7	2	7		
8	1	4	9	
9	5	6		

ones digits

Key: 5 | 8 represents 58.

Measures of Variation

The study of **quartiles** is another way to help learn about the nature and tendency of data.

Quartiles divide data that is arranged in order from least to greatest into four equal parts.

The ***median***, sometimes referred to as the Second Quartile, separates the data in half.

The ***Lower Quartile (LQ),*** sometimes referred to as the First Quartile, is the median of the first half of the data.

The ***Upper Quartile (UQ)***, sometimes referred to as the Third Quartile, is the median of the second half of the data.

The ***range*** of the data is the difference between the highest and lowest data values and is determined by subtracting these values.

The ***Interquartile Range*** is the range of the middle half of the data and is determined by subtracting the lower quartile from the upper quartile **(UQ – LQ).**

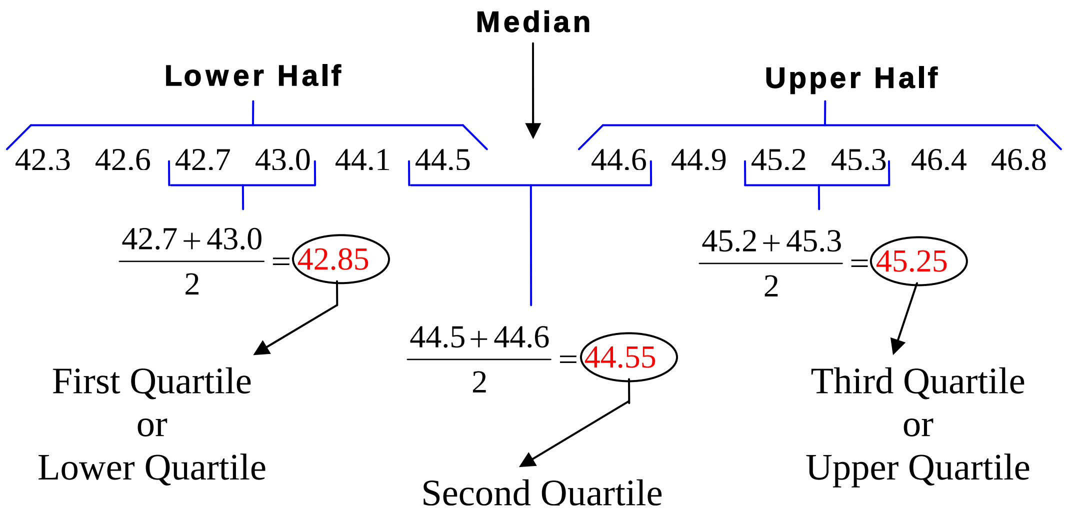
Many companies analyze data using “measures of variation” to determine the promotion and implementation of their product.

Listed below are several examples where the “measures of variation” are calculated. The data is given in order from least to greatest.

*Note: The first step in determining quartiles is to put the **data in order** from least to greatest.

Example 1: Calculate the measures of variation for the following set of data.

42.3 42.6 42.7 43.0 44.1 44.5 44.6 44.9 45.2 45.3 46.4 46.8



Range = Highest Value – Lowest Value = 46.8 – 42.3 = 4.5
Interquartile Range = UQ – LQ = 45.25 – 42.85 = 2.4

In this example there is an even number of data points in the data (12); thus, the **Second Quartile (median)** is the average of the two middle numbers. (44.55)

There is also an even number of data points in the lower half of the data (6); thus, the **First Quartile** is the average of the two middle numbers in the lower half. (42.85)

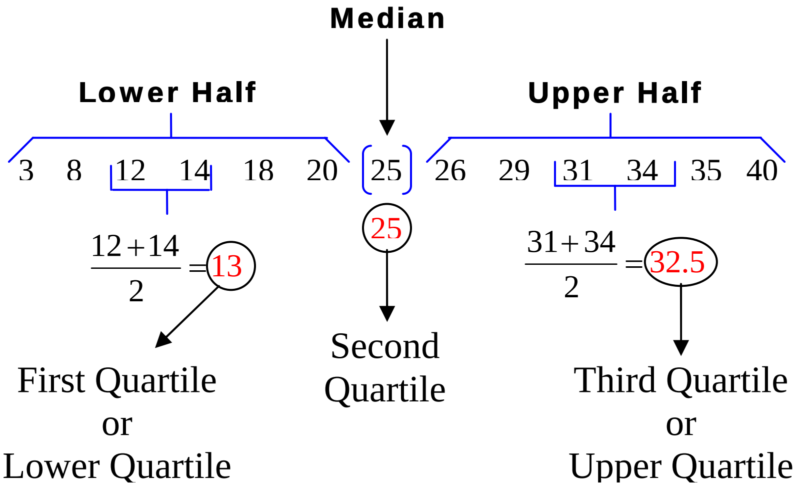
There is also an even number of data points in the upper half of the data (6); thus, the **Third Quartile** is the average of the two middle numbers in the upper half. (45.25)

The **range** is the difference between the highest and lowest data points. (4.5)

The **Interquartile Range** is the difference between the Upper Quartile (Third Quartile) and the Lower Quartile (First Quartile). (2.4)

Example 2: Calculate the measures of variation for the following set of data.

3 8 12 14 18 20 25 26 29 31 34 35 40



Range = Highest Value – Lowest Value = 40– 3 = 37
Interquartile Range = UQ – LQ = 32.5 – 13 = 19.5

In this example there is an odd number of data points (13); thus, the **Second Quartile (median)** is the middle number. (25)

There is an even number of data points in the lower half of the data (6); thus, the **First Quartile** is the average of the two middle numbers in the lower half. (13)

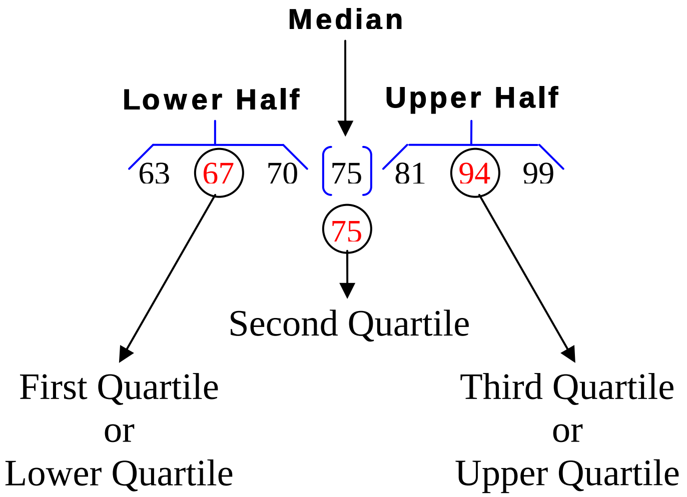
There is also an even number of data points in the upper half of the data (6); thus, the **Third Quartile** is the average of the two middle numbers in the upper half. (32.5)

The **range** is the difference between the highest and lowest data points. (37)

The **Interquartile Range** is the difference between the Upper Quartile (Third Quartile) and the Lower Quartile (First Quartile). (19.5)

Example 3: Calculate the measures of variation for the following set of data.

63 67 70 75 81 94 99



Range = Highest Value – Lowest Value = 99 – 63 = 36
Interquartile Range = UQ – LQ = 94 – 67 = 27

In this example there is an odd number of data points (7) in the data; thus, the **Second Quartile (median)** is the middle number. (75)

There is an odd number of data points (3) in the lower half of the data; thus, the **First Quartile** is the middle number in the lower half. (67)

There is also an odd number of data points (3) in the upper half of the data; thus, the **Third Quartile** is the middle number in the upper half. (94)

The **range** is the difference between the highest and lowest data points. (36)

The **Interquartile Range** is the difference between the Upper Quartile (Third Quartile) and the Lower Quartile (First Quartile). (27)

Box-and-Whiskers Plots

Box-and-whiskers plots are used to separate data into four sections. The parts will differ in length in the graph, but each part will contain one fourth of the data, with the exception of the outliers.

- **Median** - Separates the entire data set in half.
- LQ – **Lower Quartile -** Median of the lower half of the data.
- UQ – **Upper Quartile -** Median of the upper half of the data.
- **Interquartile Range** – Difference between the UQ and LQ.
- **Outliers –** Data that falls beyond the upper quartile or lower quartile and is more than 1.5 times the value of the interquartile range added to the UQ or subtracted from the LQ.
- **Lower Extreme** – Smallest piece of data excluding the outlier.
- **Upper Extreme** – Largest piece of data excluding the outlier.

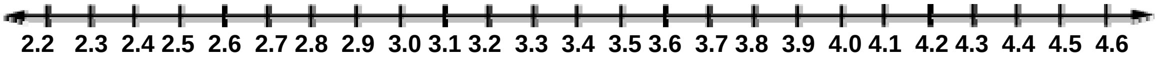
Follow the steps below to make a box-and-whiskers plot for the given data. The data represents the average monthly precipitation that occurred in Columbus, OH, in 2003. Also check the data for outliers.

Average Monthly Precipitation Columbus, OH (in inches) 2003			
2.5	2.2	2.9	3.3
3.9	4.1	4.6	3.7
2.9	2.3	3.2	2.9

Step 1: Put the data in **order** from **least to greatest**.

2.2, 2.3, 2.5, 2.9, 2.9, 2.9, 3.2, 3.3, 3.7, 3.9, 4.1, 4.6

Step 2: Draw a **number line** with a **scale** that fits the data.



Step 3: Find the **median.**

$$\frac{2.9+3.2}{2}=3.05$$

Step 4: Find the quartiles (**LQ** and **UQ**).

Lower Quartile (LQ)
Median of the **upper half**

$$\frac{2.5+2.9}{2}=2.7$$

Upper Quartile (UQ)
Median of the **lower half**

$$\frac{3.7+3.9}{2}=3.8$$

Step 5: Calculate the **interquartile range**.

Subtract the lower quartile (LQ) from the upper quartile (UQ).

$$\mathrm{UQ} - \mathrm{LQ} = \mathrm{IQ}$$

$$3.8-2.7=1.1$$

Step 6: Check for outliers.

First, check for outliers on the upper end of the data set.

- Upper limit for outliers

1. Multiply the interquartile range by 1.5.

$$1.1\times1.5=1.65$$

2. Add that number to the UQ (3.8).

$$3.8+1.65=5.45$$

3. Compare to see if any numbers in the data set are greater than this sum, 5.45.

The *highest* number in the data set is 4.6, so there are **no** outliers on the upper end of the data set. $(4.6 < 5.45)$

Now, check for outliers on the lower end of the data set.

- Lower limit for outliers

1. Multiply the interquartile range by 1.5. (Same as Step 1 above.)

$$\boxed{1.1 \times 1.5 = 1.65}$$

2. Subtract 1.65 from the LQ (2.7).

$$\boxed{2.7 - 1.65 = 1.05}$$

3. Compare to see if any numbers in the data set are lower than this difference, 1.05.

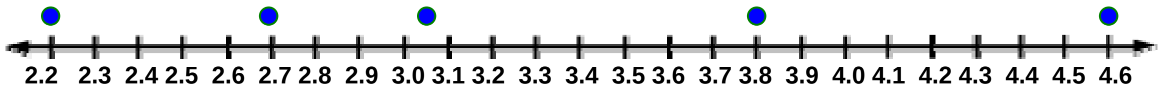
The lowest number in the data set is 2.2, so there are **no** outliers on the lower end of the data set. $(2.2 > 1.05)$

~There are no outliers in this data set.~

Step 7: Identify all critical points of the data.

Outliers: None
Lower Extreme: 2.2
Upper Extreme: 4.6
LQ: 2.7
Median: 3.05
UQ: 3.8

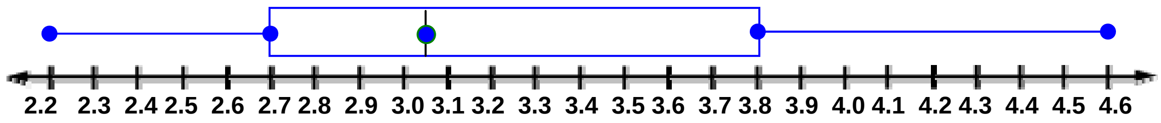
Step 8: Plot the critical points.



Step 9: Draw the **box-and-whiskers** graph.

- A rectangle (*box*) extends from the point representing the lower quartile (2.7) to the point representing the upper quartile (3.8).
- A vertical line is drawn through the point representing the median of the data set (3.05).
- A line (*lower whisker*) extends from the point representing the LQ (2.7) to the point representing the lower extreme data point (2.2).
- A line (*upper whisker*) extends from the point representing the UQ (3.8) to the point representing the upper extreme data point (4.6).
- Add a title to the graph.

Average Monthly Precipitation (in inches) for Columbus, OH, 2003



*Note: There are 12 points of data. The box-and-whiskers graph displays the data into four sections: the lower whisker, the left part of “box”, the right part of “box”, and the upper whisker. Each part contains 25% of the data.

Average Monthly Precipitation Columbus, OH (in inches) 2003			
2.5	2.2	2.9	3.3
3.9	4.1	4.6	3.7
2.9	2.3	3.2	2.9

- Three data points (2.2, 2.3, and 2.5) are located under the left “whisker”.

$$\frac{3}{12}=\frac{1}{4}=25\%\text{ of the data}$$

- Three data points (2.9, 2.9, and 2.9) are located to the left of the median under the “box”.

$$\frac{3}{12}=\frac{1}{4}=25\%\text{ of the data}$$

- Three data points (3.2, 3.3, and 3.7) are located to the right of the median under the “box”.

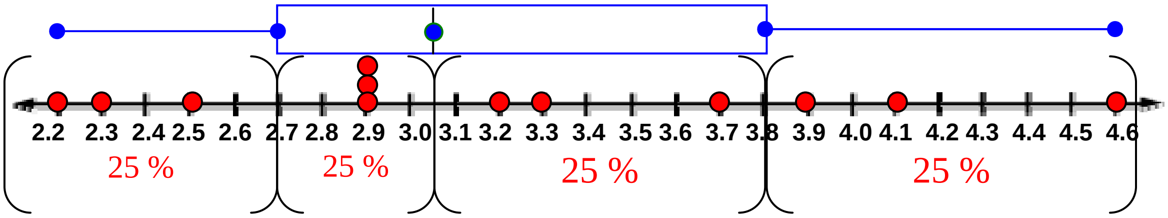
$$\frac{3}{12}=\frac{1}{4}=25\%\text{ of the data}$$

- Three data points (3.9, 4.1, and 4.6) are located under the right “whisker”.

$$\frac{3}{12}=\frac{1}{4}=25\%\text{ of the data}$$

Each part contains 25% of the data.

Average Monthly Precipitation (in inches) for Columbus, OH, 2003



Sampling

When collecting data to make predictions, it is necessary to get an **unbiased** sample selection (small group) that will be representative of the population (whole group).

Suppose Rita wanted to determine the favorite after-school activity of the students in her class by surveying a sample of the entire class.

A **biased** sample would be a survey of the members of the computer club. These members have a common interest in computers so surveying them would probably reflect a lot of computer-related activities.

An **unbiased** sample would be a survey of every fifth person listed on the class roster in alphabetical order based on his/her last name.

Let's take a closer look at the types of biased and unbiased samples that are considered for surveys.

Types of biased samples:

convenience sample – A convenience sample is a sample that includes members of a population that are easily accessed.

voluntary response sample - A voluntary response sample is a sample that involves only those people that want to participate in the sampling.

Types of unbiased samples:

simple random sample - A simple random sample is a sample where each item or person in the population is as likely to be chosen as any other.

stratified random sample – A stratified random sample is a sample in which the population is divided into similar, non-overlapping groups.

systematic random sample – A systematic random sample is a sample in which the items or people are selected according to a specific time or item interval.

Examples: Identify the type of sample described.

1. A person employed by the local mall solicits shoppers to fill in a survey about new products by offering them a lottery ticket if they take the time to complete the survey.

This is a **voluntary response sample** (biased) because the participants are choosing to take the survey (most likely because they want the chance to win a prize with the lottery ticket).

2. Every person whose telephone number ends with a 48 is contacted to find out which presidential candidate he or she favors for the next election.

This is a **systematic random sample (unbiased)** because each person surveyed is selected from a list of most of the persons in the community and based upon the condition that his/her telephone number ends with a 48 (item interval).

3. The parents, grandparents, relatives, and friends who attend the school Christmas concert are surveyed and asked if they will support the next operating school tax levy.

This is a **convenience sample (biased)** because the people surveyed were the ones that were easily accessible because they attended the school's Christmas concert.

4. Persons, ages 30 to 39, are surveyed to see which car model they prefer to purchase.

This is a **stratified random sample (unbiased)** because a select age group, ages 30 to 39, were surveyed to see which type of car they prefer. Selections from other age groups may vary considerably but if the target group for sales is this age group, then the survey is unbiased.

Samples and Predictions

A **sample** can be used to make a prediction about a population or a large group of people.

Predictions based on a sample population are *estimates*. When the sample population is representative of a larger group, the best estimates are made.

Example 1: Josh is completing a sample survey for a school project. He uses a computer and randomly selects 500 people in Belmont County, ages 25 and up, and asks if they have completed a bachelor's degree or higher? Fifty three of the 500 people respond that they have completed a bachelor's degree or higher. If the population for Belmont County is 70,226, about how many people have a bachelor's degree or higher?

What is given?

- The population in Belmont County is 70,226.
- In the sample of 500 people, 53 have completed a bachelor’s degree or higher.

What is asked?

- How many people in Belmont County, ages 25 and up, probably have a bachelor's degree or higher?

Method:

- Let *n* represent the total number of people in Belmont County who probably have a bachelor's degree or higher.

The ratio of *n* to 70,226 is equivalent to the ratio of the sample, 53 to 500. Set up the proportion and solve.

$$\frac{n}{70{,}226} = \frac{53}{500}$$ Write a proportion.

$$(n)(500) = (70{,}226)(53)$$ Cross Multiply.

$$500n = 3{,}721{,}978$$ Simplify.

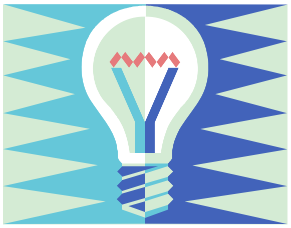
$$n \approx 7{,}444$$ Divide (3,721,978÷500).

About 7,444 people in Belmont County have a bachelor's degree or higher.

Example 2: In a manufacturing plant that produces light bulbs, a sample of 75 were selected randomly and tested. Of the 75 that were test, 3 did not perform up to standard specifications. Based on this sample, predict how many bulbs would not perform satisfactorily out of 10,000 bulbs.

What is given?

- In the sample of 75 light bulbs, 3 did not perform up to standard specifications.



What is asked?

- How many light bulbs would probably not perform satisfactorily out of 10,000 light bulbs?

Method:

- Let *n* represent the total number of light bulbs that probably would not perform satisfactorily. The ratio of *n* to 10,000 is equivalent to the ratio of 3 to 75. Set up the proportion and solve.

$\frac{n}{10{,}000} = \frac{3}{75}$ Write a proportion.

$(n)(75) = (10{,}000)(3)$ Cross Multiply.

$75n = 30{,}000$ Simplify.

$n \approx 400$ Divide (30,000÷75).

About 400 light bulbs would not perform satisfactorily.